Automatic Summarization Method for First-person-view Video Based on Object Gaze Time

Keita Hamaoka[1,*], Yasuyuki Kono[1]

[1] Kwansei Gakuin University, 2-1, Gakuen,
Sanda, Hyogo, 669-1337, Japan
{dwd11537, kono}@kwansei.ac.jp

Abstract. Several first-person lifelog videos are lengthy in duration, and often include scenes that are not useful. This can be problematic for users as it requires a considerable amount of time to watch such a video. Therefore, in this study, we propose an automatic video summarization system for first-person-view videos by employing gaze tracking and object detection, based on human gaze time on an object. Because gaze is useful for capturing a user's intention and interest, our approach summarily captures their interest and conscious focal points while watching videos. As a result of the experiment, the evaluation value of the summary video generated by the proposed system exceeded that of the summary video in which important scenes are randomly extracted. From these results, it can be said that our system is useful for rapidly watching videos, and summarizing them to reflect user interest. Our system is applicable in many fields, including behavior recognition, visual diary creation, and support for patients having memory impairment.

Keywords: Video Summarization・Gaze Tracking・Object Detection

1 Introduction

First-person-view video is a method of creating video records of daily life, sporting events, etc. It has become widespread as wearable cameras have become smaller, cheaper, and hands-free, thereby enabling a user to record their natural actions. Unfortunately, several first-person lifelog videos are lengthy in duration, and often include scenes that are not useful to viewers. This can be problematic for users as it requires a considerable amount of time to watch such a video. Several studies have been conducted with a focus on video summarization. For example, Higuchi [1] et al. summarized first-person-view video based on four cues that corresponded to the basic user actions of body movement, stillness, hand movement, and human interaction. The user

[*] Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames (first name then last name). This determines the structure of the names in the running heads and the author index. No academic titles or descriptions of academic positions should be included in the addresses. The affiliations should consist of the author's institution, town, and country.

set the importance of the four cues, and the scenes having high importance were reflected in the video after summarization. The contents of the input video were not considered, because the cues were limited to four. Our system can consider the contents of the input video which is different from previous work. In this study, we propose an automatic video summarization system for lengthy first-person-view-videos by employing gaze tracking and object detection. Because gaze tracking is useful for capturing a user's intention and interest, our approach reflects the interest areas and conscious focal points while watching a video by summarizing the video with a focus on the abovementioned information.

2 System Overview

Our system extracts a user's gazing point using a gaze tracking device, and object area using object detection function. For handling potential object detection failures, our system performs frame interpolation, i.e., it applies detection information into the preceding frame. Our system compares the distances of each object area between different frames to obtain the time-series information of each object. The object having the smallest Euclidean distance between the current and preceding frames is regarded as the same object. The object gaze time is then calculated based on the object area and the gazing point. When the gazing point is within the detected object area, count the object gaze time. If the gaze time for an arbitrary object exceeds the predefined threshold, the scene is considered important. In the generated summary video, important scenes are played back at normal speed, and others are played back at high speed.

3 Calculation of Object Gaze Time

3.1 Gazing-point Extraction

Our system extracts a user's gazing point using a gaze-tracking device. The data obtained from the device include noise. Therefore, our system smooths the data. Manu [2] et al. presented a method of smoothing gaze data by employing a weighted average. On the basis of that method, if the gazing point in the n^{th} frame is defined as Pn, the gazing point in the current frame, $P_{fixation}$, is given by Eq. (1).

$$P_{fixation} = \frac{(1P_0+2P_1+\ ...+nP_{n-1})}{(1+2+\ ...+n)}\ . \tag{1}$$

3.2 Object-area Extraction

Our system extracts the object area using YOLO v2 [3], an object detection algorithm that employs a convolutional neural network. Our system compares the distances between each object area of the current and preceding frames to obtain the time-series information of each object. If the upper-left and lower-right x and y coordinates of the object area are defined as *xlt*, *ylt*, *xrb*, and *yrb*, then the Euclidean distance between the object areas, *P* and *Q*, is given by Eq. (2).

$$d(P,Q) = \sqrt{(xlt_p - xlt_q)^2 + (ylt_p - ylt_q)^2 + (xrb_p - xrb_q)^2 + (yrb_p - yrb_q)^2}\ . \qquad \textbf{(2)}$$

An object having the smallest Euclidean distance between the current and preceding frames is regarded as the same object. The output result is shown in Fig. 1.

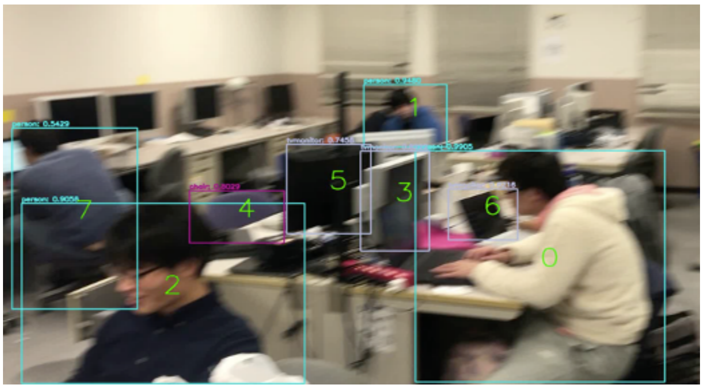


Fig. 1. Each object area is assigned a number that identifies the same object

3.3 Object-gazing-time Calculation

The object gaze time is calculated based on the object area and the gazing point. When the gazing point is within the detected object area, count the object gaze time. If the upper-left and lower-right *x* and *y* coordinates of the object area are defined as *xlt*, *ylt*, *xrb*, and *yrb*, the *x* and *y* coordinates of the object area are defined as *gazeX* and *gazeY*, respectively. The conditional expression for adding the object gaze time is given by Eq. (3).

$$\begin{cases} xlt \leq gazeX \leq xrb \\ ylt \leq gazeY \leq yrb \end{cases} . \qquad \textbf{(3)}$$

4 Summary-video Generation

4.1 Important-scene Extraction

Our system determines the importance of scenes based on the object gaze time. If the gaze time for an arbitrary object exceeds a predefined threshold, the scene is considered to be important. The user can adjust the length of the summary video by changing the predefined threshold.

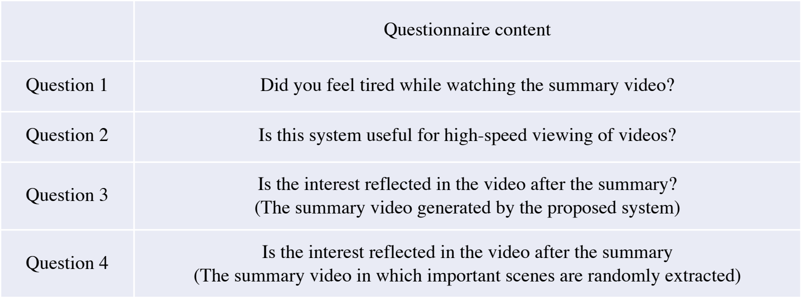
4.2 Generation of Summary Video Based on Importance of the Scenes

Based on scene importance, the first-person-view video is divided into important and non-important segments. The frame rate of important scenes is set to normal speed, and that of the non-important scenes is set to high speed. Subsequently, our system combines them.

5 Evaluation Experiment

5.1 Experimental Method

The subjects included six males and three females. They recorded a 1-h-long first-person-view video using a wearable camera and our system summarized those videos. The subjects watched both the summary video generated by the proposed system and a different summary video in which important scenes were randomly extracted. Questionnaires that used a 5-point Likert scale were then provided to the subjects. The questionnaire content is shown in Fig. 2. We also provided a free-text section for addressing the advantages and disadvantages of our system.



	Questionnaire content
Question 1	Did you feel tired while watching the summary video?
Question 2	Is this system useful for high-speed viewing of videos?
Question 3	Is the interest reflected in the video after the summary? (The summary video generated by the proposed system)
Question 4	Is the interest reflected in the video after the summary (The summary video in which important scenes are randomly extracted)

Fig. 2. Questionnaire contents

5.2 Result

Fig. 3 shows the results of the questionnaire. Fig. 4 shows the average and standard deviation of the questionnaire results. The standard deviations were calculated to two significant figures.

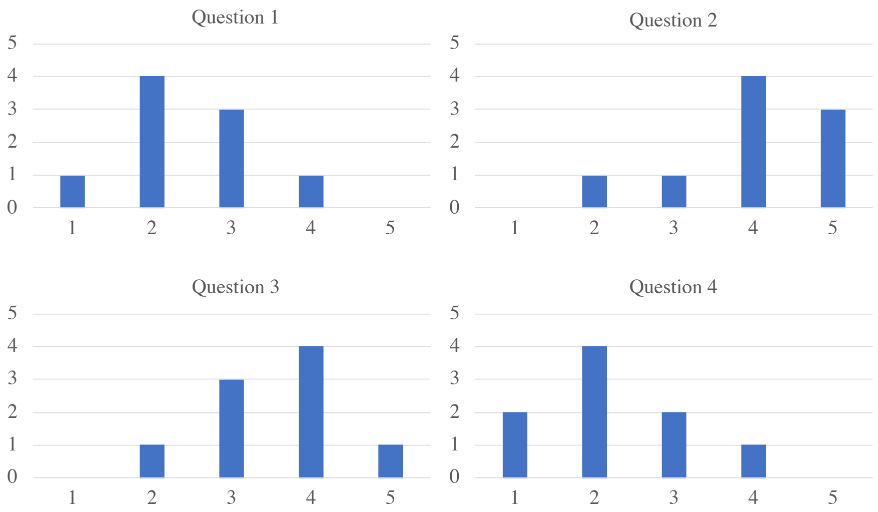
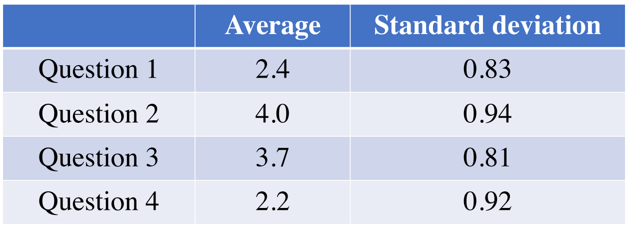


Fig. 3. Questionnaire results



	Average	Standard deviation
Question 1	2.4	0.83
Question 2	4.0	0.94
Question 3	3.7	0.81
Question 4	2.2	0.92

Fig. 4. Average and standard deviation of questionnaire results

All summary videos were less than one-third the length of the input video. A high evaluation was obtained in Question 2. The evaluation value of Question 3 exceeded that of Question 4 and considerable difference was confirmed at a significance level of less than 5%. From these results, it was found that our system accurately reflects user interests and their conscious focal points. The evaluation value of Question 1 was low. Since the frame rate of the summary video created by our system is not fixed, it is possible that the user was stressed while watching the video. These problems can be solved by excluding non-important scenes and reflecting only important scenes in the video after summarization.

6 Concluding Remarks

We proposed an automatic summarization system for first-person-view videos based on object gaze time. From the results of an evaluation experiment, we observed that our system is useful for rapidly watching videos, and summarizing them to accurately reflect user interests. A user opinion was provided in the free-text area that stated, “the scene when I was losing interest was regarded as an important scene and it was reflected in the summary video." Because this system sets important scenes based on the object gaze time, it is not possible to consider the user's mental state. We believe that these problems can be resolved by employing information regarding a user's pupil diameter as there is a correlation between pupil diameter and wakefulness. Thus, our proposed system will be able to accurately reflect mental information by considering the change in pupil diameter. To achieve this, the challenge will be to find the threshold of the amount of change in the pupil diameter during a scene for which the user has a high degree of interest.

References

1. Higuchi, K., Yonetani, R., Sato, Y.: EgoScanning: quickly scanning first-person videos with egocentric elastic timelines. In: 2017 CHI Conf. Human Factors in Computing Systems, pp. 6536–6546 (2017)
2. Manu, K., Klingner, J., Puranik, R., Winograd, T.: Improving the accuracy of gaze input for interaction, In: 2008 Symposium on Eye Tracking Research & Applications, pp.65–68 (2008)
3. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger, In: 2017 IEEE Conf. Computer Vision and Pattern Recognition, pp. 7263–7271. IEEE Press, New York (2017)